


RESEARCH

Open Access



# Train rolling stock video segmentation and classification for bogie part inspection automation: a deep learning approach

Kaja Krishnamohan<sup>1</sup>, Ch. Raghava Prasad<sup>1</sup> and P. V. V. Kishore<sup>2\*</sup> 

\*Correspondence:  
pvvkishore@kluniversity.in;  
pvvkishore@gmail.com

<sup>1</sup> Department of Electronics and Communications Engineering, K.L. University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh 522 502, India

<sup>2</sup> Image Speech and Signal Processing Research Group, Department of Electronics and Communications Engineering, Biomechanics and Vision Computing Research Center, K.L. University, Green Fields, Vaddeswaram, Guntur DT, Andhra Pradesh 522 502, India

## Abstract

Train rolling stock examination (TRSE) is a physical procedure for inspecting the bogie parts during transit at a little over 30 kmph. Currently, this process is manually performed across many railway networks across the world. This work proposes to automate the process of TRSE using artificial intelligence techniques. The previous works have proposed active contour-based models for the segmentation of bogie parts. Though accurate, the models require manual intervention and are found to be iterative making them unsuitable for real-time operations. In this work, we propose a segmentation model followed by a deep learning classifier that can accurately increase the deployability of such systems in real time. We apply the UNet model for the segmentation of bogie parts which are further classified using an attention-based convolutional neural network (CNN) classifier. In this work, we propose a shape deformable attention model to identify shape variations occurring in the video sequence due to viewpoint changes during the train movement. The TRSNet is trained and tested on the high-speed train bogie videos captured across four different trains. The results of the experimentation have been shown to improve the recognition accuracy of the proposed system by 6% over the state-of-the-art classifiers previously developed for TRSE.

**Keywords:** Deep learning, Train rolling stock, Automation, Convolutional neural networks, UNet

## Introduction

The mass public transport in the Indian subcontinent is commuted using railways. Indian Railways is the largest network covering 68 km with 13 K passenger trains running across this rail network. Indian Railways serve around 23 million passengers every day. Moreover, more than 50% of these trains run for more than 500 km at a time. Consequently, operational research shows that the health of the train is directly proportional to the safety of the train which in turn affects the ride quality of the passengers. To ensure enhanced safety of the 100-crore train and its passengers, the most widely practiced inspection mechanism during the train movement is called Train Rolling Stock Examination (TRSE).

TRSE can be considered as an operational maintenance service to examine the bogie parts on which the train moves. This examination is performed when the train is moving at just over 30 kmph. The Indian Railways operational manual (<https://rdso.indianrailways.gov.in/works/uploads/File/Draft%20Handbook%20on%20Integrated%20Rolling%20Stock%20Depot.pdf>) on TRSE provides details of the various checklists on the performance of the bogie parts that are visually observable during the process. A few instances from the manual are as follows: axle box leaks, suspension movements, hanging parts, brake shoe functionality, breakages in parts, missing screws, and flat tires. The entire operation is performed by a three-man crew, one on each side of the train and the other one recording the defects. The defects are then classified into immediate action-required maintenance jobs and pit-stop maintenance jobs.

The pit stop maintenance jobs can be executed during the overall maintenance of the train at the designated destination in a railway maintenance yard. Contrastingly, the immediate jobs are handled when the train halts in the following station. The entire TRSE is fully proven that it has been in practice for 100 years. However, the failure of this process has also caused accidents and loss of life bringing a great financial burden on the railway operations. The biggest reason for failure has been analyzed as manual monitoring and no mechanized support system. We are the first in the subcontinent to provide a visual support system for TRSE to assist the monitoring engineers [1–7].

Our previous models focused on bogie part shape extraction from the videos of train undercarriage. The works focused on developing active contour models with multiple shapes prior to knowledge-based inferences. These models did a great job of preserving the shape of the segmented bogie parts. Despite their success in bogie part segmentation, active contours have an underlying computational complexity when it comes to high-resolution video data [8, 9]. The bottleneck in applying knowledge-based active contours is attributed to its iterative model. These are energy-based models that propagate a differential contour by optimizing the contour regions with respect to the divergence of the image. Hence, real-time implementation of these methods has been next to impossible.

Deep learning approaches have shown to have deployable capabilities and were being used for many video object recognition [10–13] and analytics [14–16] applications. Our previous model used a modified Yolo V2 architecture for the bogie part identification process on video data [17]. However, the challenge was to annotate the maximum possible part variations across different training video datasets. Moreover, the architecture Yolo V2 was modified to make the attention mechanism more stringent on the bogie parts to detect their presence across the entire video sequence. This process made the training loads heavy and the computation process cumbersome during training even though the testing was simpler. Despite its good performance, it lacked two major objectives of TRSE: (1) the bogie part shape extraction or descriptor for determining the health of the parts during transit and (2) the deformation in the bogie parts in the consecutive video frames as the train moves horizontally with respect to the camera angle.

The above two objectives along with the core objective of segmenting bogie parts and recognizing them with high accuracy will be successfully resolved using the proposed deep learning framework. Our proposed deep network has two frameworks: (1) the segmentation module and (2) the attention-based classifier. The traditional object detection

deep learning models apply annotations on all the bogie video frames to supervise the recognition process with good accuracies. Moreover, these methods do not detect the structure or shape similar to the bogie parts for maintaining structural durability during transit. Consequently, the segmentation results at multiple stages were applied as attention to the classifier for achieving higher resolution rates. The attention mechanism followed is inception from transformers for speech recognition applications which has been incepted into our proposed model.

Specifically, this part of the section presents the past research on technological developments on the road to automate TRSE. The entire section has three main ingredients. The first one is related to methods developed in general towards the solution of rolling stock examination. Secondly, the computer vision-based models were applied to video data for the bogie shape segmentation process. Thirdly, the advancement of deep learning approaches and their applications in object detection.

### **TRSE technology-based approaches**

Indian Railways (IR) is the largest rail network on the planet which carries 10 million passengers every day. The primary challenge is to keep the trains away from accidents. Consequently, this is the job of railway maintenance engineers and researchers to develop new technologies to assist human resources. Currently, advanced technologies are being applied for efficient locomotive production [17], high reliability coach design [18], electronic signaling system [19], a Global Positioning System (GPS)-based train tracking [20], and ultrasound reflectors for track anomaly detection [21]. Apart from the above technologies for uninterrupted operations, the important aspect of train and passengers safety is rolling stock examination [22]. TRSE is performed manually across all rail companies using 3 humans and specifically using a rolling pit. The success rate of Manual TRSE was found to be around 99% for an entire decade of operations in Indian Railways (<https://rdso.indianrailways.gov.in/works/uploads/File/Draft%20Handbook%20on%20Integrated%20Rolling%20Stock%20Depot.pdf>). However, that 1% failure rate had damaged millions of dollars in the economy and thousands of lives. That is why automating this TRSE becomes an increasingly important problem to find a solution. The objective of this work is to transform the manual rolling stock examination into an automated or semi-automated system to assist railway engineers. Indian Railways are testing a prototype model named KRATES (Konkan Railways Automated Train Examination System) ([https://konkanrailway.com/uploads/editor\\_images/1551089341\\_ATES%20web%202022022019.pdf](https://konkanrailway.com/uploads/editor_images/1551089341_ATES%20web%202022022019.pdf)). The system has many sensors to measure temperature, pressure, acceleration, brake shoe functionality, and a camera. The purpose of the camera in the KRATES module is about remote visual monitoring rather than real-time predictions on the bogie parts.

According to the Indian Railways rolling stock manual, the following primary checks are needed for a train to consider fit to reach the next destination without any accidents.

1. Hanging bogie parts
2. Broken bogie parts
3. Dragging bogie parts
4. External agent in the bogie parts

5. Flat wheels
6. Missing bogie parts

The above parameters are all visually observable and are biologically compared with the training models for evaluation. This results in documentary evidence that provides an insight into the behavior of the bogie parts during transit. The objective of this work is to transform the above visually identifiable problems into computer vision-based models for automated TRSE. Currently, most rail network companies operate manually due to unavailability of technology or research resources for finding a commercially viable solution. However, video-based bogie part retrieval models have been developed in the past with considerable research impact in the field of computer vision.

#### **Computer vision algorithms for TRSE**

Initially, the work in [23] has been a starter to unfold the deeper connection between transport automation technologies and their ability to prevent accidents. Inspired by this, some of our previous works have been built on the basis of computer vision. Most of the research works on train safety are different from ours as they are focused on rails and ballast monitoring using computer vision. The work in [23] shows the first use of a camera to record bogie parts and extract them using image mosaicking. 3D imaging models were designed and developed for inspecting train tires and the surrounding ballast using simple 3D correlations [24]. However, the dual cameras were not grounded but are mounted on the train. The movement of the train captures the rails and the ballast with predefined displacements. The parallel projection model along with the 3D digital image correlations will identify anomalies in tracks and ballast. However, the biggest drawback is the train movement at high speeds will make the video data blurry making faulty measurements. Another dual camera model with multi modal recordings in RGB and infrared frequencies has been applied for the bogie part identification process [25]. The method developed uses a panoramic viewing model to compare RGB and infrared images to locate bogie parts. The infrared camera has been in place to identify heating bogie parts such as axle box, brake shoes, joints, and high friction contacts. The two cameras were used to detect hot and cold parts simultaneously. However, motion blurring in RGB video data and high ambient temperatures make the detection process ambiguous. The next work shifts to a pit hole camera system placed inside a trench dug under the tracks to capture brake shoes [26]. The keyframes with break panels are extracted, and curve fitting models were applied to segment the break portions and identify defects. Even though the results were prominent in the brake shoes functionality identification process, the actual implementation of the project poses a bottleneck both commercially and structurally. Currently, TGV of France and bullet trains of Japan use train-mounted cameras to monitor tracks and ballast. The monitoring is executed manually, and no processing algorithms were reported in the published patent [27]. This is due to the fact that the video sequences captured from cameras mounted on high-speed trains are subjected to unpredictable vibrations which generate noisy video data for processing. Interestingly, the camera sensor on the ground has shown to achieve maximally effective train bogie video data for monitoring than the system mounted on the train. One such system was developed with lights and antiglare techniques for capturing high

contrast video frames of rolling stock [28]. Actually, this work has been the basis for automating TRSE. However, this work does not highlight anything about the algorithms for bogie part identification. Another work that has drawn parallels with the above has demonstrated the use of focus lights on the undercarriage to video capture the bogies [29]. Additionally, this work applies basic image processing models to extract the edges of bogie parts in order to identify them. However, the techniques described were not able to represent the overlapping boundaries of bogie parts in the video frames. Moreover, the blurring induced by the moving train has made the edge detection process difficult for part identification. Recently, 3D modelling of contact bogie parts and wheel surfaces has been shown to achieve good results for the detection of defects [30, 31]. The biggest problem with 3D modelled image data is their powerful graphics processing requirements. The powerful graphics make these techniques incompatible with real-time processing.

The two biggest drawbacks of the above models were their inability to segment bogie parts effectively, and the video data was noisy due to recording of train movement at 30 frames per second shutter speed. These two bottlenecks were efficiently handled by our previous models for TRSE [1]. To fight blurring, the recording is done by using a high-speed wide-angle sports action visual sensor at 240 fps, the effectively exceptionally high-quality bogie frames. Secondly, the segmentation problems were addressed using active contour (AC) models with shape prior knowledge of the bogie parts [2, 3]. These shape-based active contours with local information [5] have presented a 99% accuracy in preserving the extracted bogie part shapes from the output of the models. Moreover, the work in [4] shows an upgraded touching Boundary segmentation algorithm for collectively extracting bogie parts from the video frames. This model has generated interest due to the fact that the bogie parts are indeed overlapping as they support each other to tightly hold the entire structure as a single unit. The above AC-based models have performed well in segmenting the bogie parts effectively. Despite their success in bogie part segmentation, the AC models are iterative and are not suitable for real-time implementation of TRSE. Apart from the above, the TRSE automation algorithms lack adaptability, scalability, and reliability to transform the results into real-time production models. Consequently, these gaps in current research methodologies have motivated us to perceive the real-time implementable models for automating TRSE. Hence, the deep learning approaches were leveraged to build and deploy automated TRSE systems for generating actionable intelligence for assisting rail companies.

### **Deep learning approaches for automating TRSE**

The implementable learning (DL) approaches have been in operation since 2012 with the creation of AlexNet in the ILSVRC ImageNet challenge [32]. The then AlexNet has been trained on 3 GB GPUs from Nvidia using parallel processing. After that, many highly accurate and reliable models have outperformed the AlexNet. They are inception V1 [33], VGG-16 [34], ResNet [35], SeNet [36], and PNASNET [37]. These models have been shown to achieve a very high rate of accuracy in image classification tasks over the years. These base models were updated to detect objects in images and video sequences, and one such model that has performed consistently on multiple test sets is the Yolo (You Only Look Once) architecture [38]. In our previous work [5], the second version

of Yolo is modified for the extraction of bogie parts on the video sequences. The model was able to detect most of the bogie objects except for the places where the part deformation is more than 50% of the actual trained part. The biggest challenge in implementation is attributed to the annotations of bogie parts from video sequences along with the bounding box information on which the Yolo model is trained. Though the model has recorded an 84.98% accuracy in correctly identifying the bogie part on a moving train video sequence, it failed to identify bogie parts with high confidence scores for the slightest deformation in the objects occurring due to viewpoint variations. Moreover, to compensate for the object deformations, the model has been trained on a large set of frames in the video sequence. Hence, it becomes extremely important to learn the object deformations for the segmentation process. In deep learning, the segmentation process has been applied through an architecture broadly called as hourglass model [39]. Then, the upgrades with some minor modifications have reported betterment in segmentation results, though their basic structure matches the hourglass model. The most popular and powerful variants of the hourglass are UNet [40], VNet [41], SegNet [42], and Auto Encoders [43]. The backbone network architectures in these segmentation modules can be any of the state-of-the-art network architectures such as VGG-16, Resnet-34, and Inception Net. Once the segmentation processes are learned by the network using a very small dataset of bogie parts, the next step is classification. Generally, instances have shown that the segmented output is inputted along with the original video frame into the classifier for recognition. The RGB input is multiplied with the segmented bogie parts and passed to the classification module designed using the standard networks similar to that of the backbone segmentation network [44, 45]. Unfortunately, doing the multiplicative attention will instigate the user to segment all the bogie parts in all the video frames for maximally correct classification. Instead of performing the traditional multiplicative fusion between the RGB video frames and the segmented objects, this work offers a solution incepted from the model of natural language processing called multi-head attention [46]. Similar methods were proposed in the automation of construction durability testing such as identifying payment cracks using capsule net segmentation [47] and PCGANs [48].

Finally, the proposed model brings a novel methodology for real-time implementable automated TRSE powered by computer vision, artificial intelligence, and video analytics. The next section focuses on developing a detailed elaboration of the methods applied for automating TRSE with deep learning.

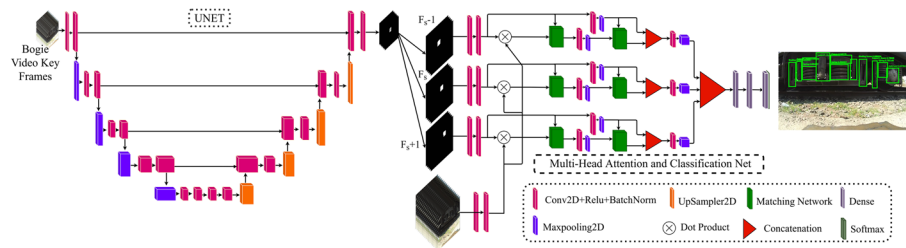
## Methods

The highly accurate and accepted attention model in speech processing is the multi-head attention model. The multi head attention model is capable of providing attention to a particular set of words during training. Similarly, the moving train induces motion artifacts such as bogie part shape deformation due to viewpoint variations on the fixed camera positioning. Moreover, the camera outputs 240 fps video sequences with a lensed angle of 54°. The method proposed in this work has a segmentation network followed by multi head attention-based bogie parts classifier. The entire model is called deep bogie part inspector (DBPI) which has a segmentation module in the back end and an attention-based classifier model in the backend. The primary network in the DBPI model is





**Fig. 1** Video frames of train rolling stock of a moving train at 30 kmph



**Fig. 2** Deep bogie part inspector architecture with two models: the first model is UNet used for segmentation of bogie parts, and the second is the multi-head attention-based classification network

built on the UNET architecture backbone. Subsequently, the secondary network is a classifier model built multi-head-head attention network. In our proposed multi head attention network, we have three streams that are fed with three consecutive segmented bogie parts in keyframes. The proposed method has been shown to improve the recognition accuracies by 6% over the existing object detection deep networks.

The proposed work has three core objectives: (1) to segment the bogie parts from key video frames of train undercarriage by learning from deformations in shapes and spatial locations, (2) to apply the segmented bogie parts to find an attention matrix that will contribute to the faster identification of bogie parts in continuous video sequences, and (3) to design a multi head attention-based architecture for the classification of bogie parts irrespective of their shape and spatial location in the entire video sequence.

Finally, the proposed model will also give a bogie part assert score (BPAS) that can help the human TRSE inspector to make decisions for timely maintenance and thereby increasing passenger safety. Figure 1 shows the video frames of the bogies on Indian Railway coaches.

Our proposed deep bogie part inspector (DBPI) for TRSE is different from the existing models in three different ways: (1) multi head attention network in the classifier will learn from a minimalistic dataset making the training process faster, (2) it offers higher bogie part classification accuracies across the entire range of video sequences, and (3) the model generates actionable intelligence for the maintenance engineers to predict the durability of the bogie parts during the train running cycle.

The deep bogie part inspector is an ensemble of two learning networks as stated in the introduction. The architecture of the proposed DBPI is shown in Fig. 2. The first network is based on UNet architecture to segment the bogie parts from the video sequences. Followed by the UNet is the classification network that identifies a bogie part and checks its durability for the onward train journey towards the destination. Specifically, the classifier is built on multi head attention mechanism where the segmented output of each part is applied to determine the attention of the part in the video sequences. Furthermore, the

matching networks were designed to establish a semantic correspondence between the parts and the original video frames. This process allows the network to match the correct position of the bogie part from multiple frames resulting in the correct match from a few sampled segments. This has enabled the network to learn from a few frames of segmented bogie parts rather than the bogie parts in all the frames. Finally, the extracted features are concatenated locally first and then globally before being learned by the fully connected neural nets. The last layer is a Softmax that predicts the correct bogie part from the input video frames.

### The bogie segmentation module — the B-UNet

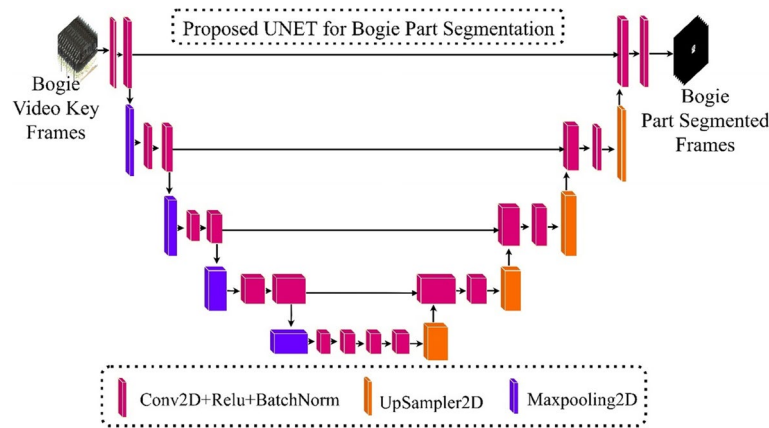
The bogie-U-shaped convolutional neural network (B-UNet) is a segmentation module for separating the bogie parts individually from the video frames is shown in Fig. 2. Given a complete set of bogie video frames  $V(x, y, 3, t) \forall (x, y, 3) \in R^2$ , where  $t$  is the frame number, the objective of B-UNet is to segment the bogie parts  $S_b^k(x, y, k)$ , where  $k$  is the pointer to the key frames. The key frames are important as the video is captured with 240 fps shutter speed; the number of frames within a second is equal to 240. The change across 240 frames is less than noticeable by the artificial visual sensor, and hence, key frames are extracted. Since all the bogie video frames have similar pixel densities the feature-based key frames extraction models using histogram of oriented gradients (HOG) features with K-means clustering had little impact on the outcome. However, the entropy-based method [49] has shown a good deal of variation in pixels across the video frames. The frame entropy is computed as follows:

$$E(f) = - \sum_j p_f(j) \times \log(p_f(j)) \quad (1)$$

The entropy  $E$  of frames  $f$  is a 2D space between  $E$  and  $f$ . The 2D entropy space could offer local maximum and minimum points from which local extreme points are extracted. These extracted points are the frames representing the key frames. The resulting key frames of bogies are given by  $V(x, y, 3, k)$  where  $k=1, 2, \dots, K$  and  $(x, y, 3)$  is the pixel locations in 3 dimensions.

Now, we redefine the problem of segmentation as, given a set of key bogie frames  $V(x, y, 3, k) \forall k=1$  to  $K$ , design a UNet model to learn the bogie parts binary models for segmentation  $S_b^k(x, y, k)$ . The parameter  $b$  gives the number of bogie parts or bogie part index. The architecture of the UNet has been incepted from [47]. The model takes an input frame size equal to  $240 \times 424 \times 3$ . The network in Fig. 3 has 8 convolution layers with 16,32,64,128 filters per two consecutive layers in both the compressive and expansive paths as shown in the UNET part of Fig. 2. All layers have  $3 \times 3$  unpadded convolutions with ReLu activation functions followed by a maximum pooling of  $2 \times 2$ , which halves the frame resolution to the forward layers. Subsequent down-sampling steps will see a doubling of the number of filters in both arms of the UNet. There are no fully connected layers in the end of the downsampler block. Subsequently, upsampler blocks add pixels in  $2 \times 2$  up-convolutions that cut down the number of filter channels to half of the corresponding counterpart in the downsampling level. This results in the loss of feature channels during upsampling making it difficult to generate a segmentation mask with these small feature maps. This loss in feature maps is compensated with the help of skip





**Fig. 3** The proposed UNet bogie part segmentation module

connections that apprehend the cropped feature maps from the corresponding compressor blocks after the up convolutions in the expander blocks at the same level. The cropping of the encoder or compressor block features is necessary to ensure that uniform dimensionality for concatenation with the expander feature maps. These concatenated feature maps are further learned using two  $3 \times 3$  convolutional layers followed by a non-linear ReLu. Finally, a  $1 \times 1$  convolutional layer maps each of the 16-component feature vectors into the required classes. Our B-UNet has only 16 components against the 64 in the original UNet architecture. This is because the segmentation stroke of the bogie part in the entire video frame is small compared to the entire spatial resolution of the frame itself. After multiple experiments, the 16-channel filter is perfect for bogie part segmentation and is computationally faster than the traditional UNet model.

The bogie segmentation has a large background region when compared to the spatial occupancy of the part in the video frame. This has resulted in a dominating loss with respect to the background during the training process falling into the local minimum frequently. Hence, we propose to apply the solution in [62], which addresses the foreground-background pixel imbalance in the rolling stock video frames. We applied the two traditional loss functions during the training process. They are a type of binary cross-entropy (BCE) called focal loss (FL) and dice loss (DL). The BCE is given by

$$FL(G_T, p) = \begin{cases} - \sum_{i=1}^{np_0} \alpha (1 - \alpha)^\gamma \log(p), & \text{if } G_T = 1 \\ - \sum_{i=1}^{np_1} \alpha (1 - \alpha)^\gamma p^\gamma \log(1 - p), & \text{otherwise} \end{cases} \quad (2)$$

where  $G_T$  is the ground truth in the pixel range  $\{0, 1\}$  and  $p \in [0, 1]$  is the probabilities of foreground and background predicted by the model.  $\{np_0, np_1\}$  are classes that represent background class with 0th values and foreground with 1 value. The values of  $\alpha \in (0, 1]$  and  $\gamma \in [0, 5]$  are adjustable hyperparameters. For B-UNet, we selected  $\alpha = 0.5$  and  $\gamma = 1$  across all datasets.

The second loss used was dice loss (DL) which is a regular in segmentation problems using deep learning models. Dice loss solves the problem of imbalance between foreground and background pixels using the segmentation evaluation index between the

predicted segmentation mask and ground truth annotated masks. The DL is formulated as follows:

$$DL(p, G_T) = 1 - \frac{2 \sum_{i=1}^{np} p_i G_{T_i} + \delta}{\sum_{i=1}^{np} p_i^2 + \sum_{i=1}^{np} G_{T_i}^2 + \delta} \quad (3)$$

The parameter  $\delta \in [0, 1]$  is a preventive measure to avoid a divide by 0 instances during training. The two losses were used simultaneously for backpropagating through the network for weight modifications. However, the combination of the proposed loss is considered as an average over the entire pixel range defined as follows:

$$S_L = \frac{FL(p, G_T)}{np} + DL(p, G_T) \quad (4)$$

The B-UNet segmentation network is trained on  $K$  key frames to extract  $b_p$  bogie parts from  $B \in [1, b]$  bogies and  $P \in [1, p]$  parts. Testing is initiated on the sequences of bogie parts that were not previously seen by the B-UNet of different trains. The obtained parts are now applied as inputs to the classifier to identify the bogie part correctly and provide the necessary analysis.

#### B-UNet implementation

The original frame size from the high-speed camera sensor was  $1280 \times 1918$  at 240 fps. The sensor records 240 frames per second, and in a 1-min video, we have around  $240 \times 60 = 14,400$  frames per minute. Our dataset consists of passenger trains from the Indian subcontinent which are having an average of 20 coaches per train. The camera sensor's average recording of a train happened for around 1.05 to 1.42 min. All the above values are computed based on the video contents in our dataset. The average number of frames in each training class was found to be around 15,456 frames per train. Using the entropy-based formulation, the key frame extractor will assemble only frames with maximally occupied bogie parts. The number of frames per bogie is around 0.2% of the total frames, which is 30 frames/bogie. There will be two bogies per coach per side, and for a 20-coach train, there will be 40 bogies. Finally, the training set for bogie part segmentation consists of  $30 \times 40 = 1200$  video frames. From these 1200 training bogie video frames, we train only for 8 bogies with 18 parts. This is because the bogie parts are fairly constant over the entire train; it is unnecessary to use all bogie frames for training. A number 8 also guarantees good data augmentation for training apart from others such as rotation, scaling, zooming, and flipping horizontally and vertically in our model. Finally, the training set has 320 frames in 100 different augmentations per frame. The total dataset for B-UNet will have 32 K video frames and 32 K ground truth labels with around 1778 parts per label. The filter kernels are initialized using the zero mean Gaussian centered around unit variance. A batch normalization layer is added after each convolution layer to induce stability of the process. The hyperparameters in the loss function are selected as discussed in the previous section for all the bogie videos through experimentation. The optimizer is Adam with a learning rate of 0.00001 and a momentum factor of 0.02. There is no decay in the learning rate as the error reaches a minimum value. All these

methods are unchanged across all datasets and on other models used for comparisons. All the models were implemented on NVIDIA GTX1070i attached to 16GB memory. The epochs hyperparameter is set to 100 for all models.

The testing is performed on the full bogie video sequence without key frame extraction for segmenting the bogie parts. The segmented bogie parts are now arranged in chronological triplet order of current  $f_c$ , previous frame  $f_{c-1}$ , and next frame  $f_{c+1}$  for each bogie part. These three groups of segmented video frames form the input to the classifier which is built on the multi head attention model.

### The bogie classification module: B-MHAC

The B-MHAC (bogie–multi head attention classifier) is a combination of an attention grabber network and the dense network classifier with Softmax activation. The outputs of the  $1 \times 1$  convolutions in the B-UNet are segmented bogie parts that are separated into multiple classes manually. Given the bogie classes  $C_b$  with their segmented bogie parts from  $b = 1$  to  $B$  at 30 time steps and the raw bogie RGB video frames  $V(x, y, 3, t) \in R^2$ , the objective of B-MHAC is to learn the distinct bogie part features using the multi head attention framework as an object placeholder in the video frames. The right side of Fig. 2 has 4 streams of convolutional layers with three of them forming the basis for attention on the raw video frames in the 4th stream. The convolutional layers in each stream will accept an input of size  $240 \times 424 \times 3$ , which is interpolated to match the segmented outputs to  $240 \times 424$ , which will be operated upon by 32 filters of size  $3 \times 3$  with stride 1. These linear layers are nonlinearized with ReLu activations and passed through batch normalization to train the layers more independently.

Given the features from the four streams with sizes  $W \times H \times F|_t^i$  in the  $l^{th}$  layer of  $i^{th}$  the feature matrix, the first goal is to multiply the segments at different time scales with the incoming bogie RGB frame independently in the upper 3 streams of the multi head network. The output of the multiplication is  $M_s^l(i) \in R^2$  where  $s = 1, 2, 3$  is the stream number. In order to obtain the relationship between the original masked object  $M_s^l(i)$  and the segmented bogie part features  $f_s^l(i)$ , we apply a feature matching block as shown in Fig. 4. Here  $l$  gives the layers, and  $i$  represented feature positions.

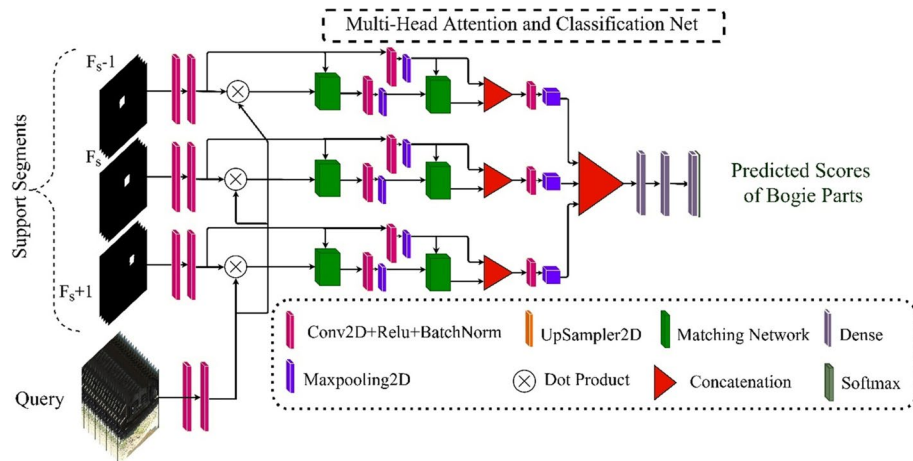
Let  $f_M \subset M_s^l(i)$  and  $f_B \subset f_s^l(i)$  be the features of query masked object and the support bogie parts of size  $W \times H \times C$ , respectively. Primarily, these two features are mapped to a spaces  $\Theta$  and  $\Phi$  to obtain  $\Theta(f_M)$  and  $\Phi(f_B)$ , respectively. Subsequently, the 3D matrices are reshaped to  $WH \times C$  which transforms into a spatial attention maps using the formulation below:

$$h(f_M, f_B) = \underset{rows}{\text{soft max}} \left( \Theta(f_M) \Phi(f_B)^T \right) \quad (5)$$

Meanwhile, the output  $h(f_M, f_B)$  is multiplied with the features in support spaces  $\Phi(f_B)$  into an intermediate space  $g(f_M, f_B)$  formulated as follows:

$$g(f_M, f_B) = h(f_M, f_B) \times w(f_B) \quad (6)$$

The  $w_j(f_B)$  are the features of the bogie parts at  $j^{th}$  position in the network. This ensures that the features that are relevant to the query image are retained and that which are irrelevant are discarded.



**Fig. 4** The proposed cascaded feature matching module with multi head self-attentions for accurate tracking of bogie part position and identification

Finally, the output of matching network  $g(f_M, f_B)$  is reshaped to that of the original query features and is concatenated with them by applying a  $\delta$  weighing rule. The formulation is computed as follows:

$$F_I = \delta \times g(f_M, f_B) + (1 - \delta)M(i) \quad (7)$$

The  $\delta$  value is a hyperparameter which will be decided based on the experimentation and the pixel density of each of the bogie objects. Finally, the integrated features  $F_I$  from each of the bogie classes are applied to a two-stage dense network with Softmax activation for classification. Though the process is computationally expensive, it has shown to recognize deforming shapes of bogie objects during the movement of the train. Accordingly, we test the performance of the proposed method through experimentation and validation on the train rolling stock dataset.

## Experimentation

### TRSE datasets

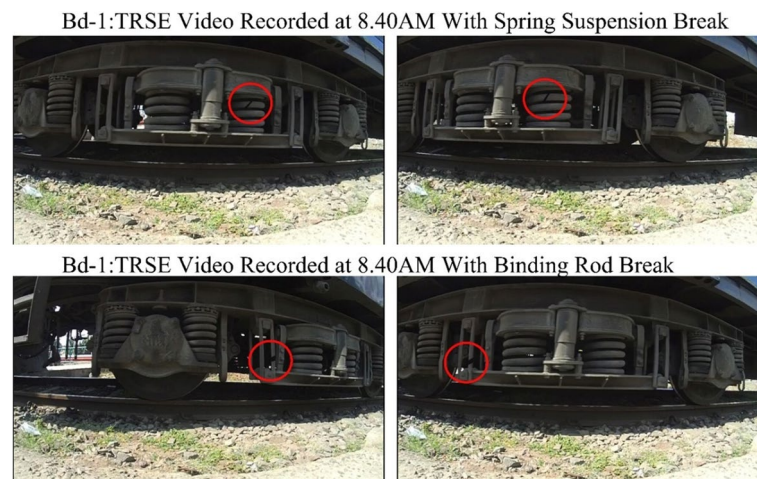
The datasets used in this work are shown in Fig. 5. A more detailed view of the capturing mechanism and sensor used is given in our earlier works [33]. The train rolling stock examination (TRSE) bogie videos are captured at different time stamps during the day as shown in Fig. 5. Each of these videos were captured at 240 frames per second when the train was moving at a little over 30 kmph at 1080P resolution. Each of the video datasets has more than 21,000 frames.

Since it was difficult to find defective bogies within a short period in real time, we simulated the defects found regularly on bogie parts using photoshop and reinduced those frames back into the original video sequence. Figure 6 shows two such defects on spring suspension and binding rods.

The objective of the experimentation is to identify the following bogie parts in the video sequence as shown in Fig. 7. Altogether, there are 16 bogie parts that should be monitored during TRSE as per the Indian railway rolling stock examination manual. The numbering will be part of the class names as there are multiple parts with the same



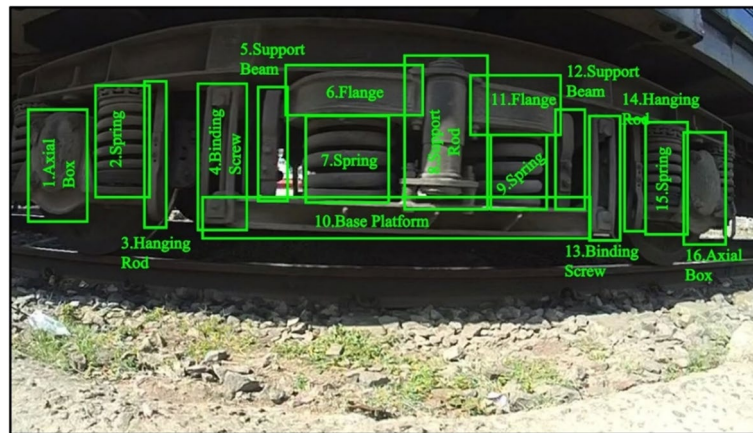
**Fig. 5** Datasets for TRSE used for experimentation



**Fig. 6** Defective bogie parts induced into the original video frames through photoshop

name. A total of five parameters were used to judge the performance of the algorithms qualitatively along with the visual validation on the test view frames. They are intersection-over-union (IoU), mean average precision (mAP), mean False Identification (mFI),





**Fig. 7** Bogie parts being identified through the proposed algorithm

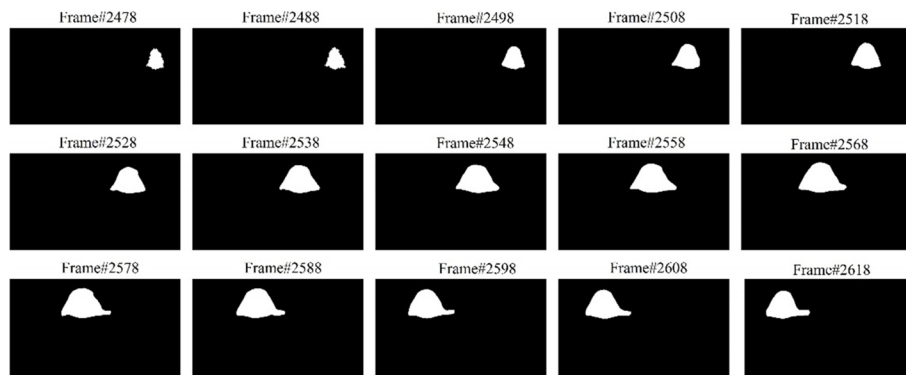
and mean non-identification (mNI). The IoU is generally used for understanding the performance of the UNet segmentation module and its role in the judgment of the classifier. The range of IoU is between 0 and 1, with the latter being the desired value for a good segmentation algorithm. Similarly, the mAP gives the precision with which the classifier identifies the given bogie object. The mFI is a parameter that indicates the false identification of a bogie object, and mNI gives the inability of the classifier to identify the bogie object.

#### ***Training and testing the B-MHAC***

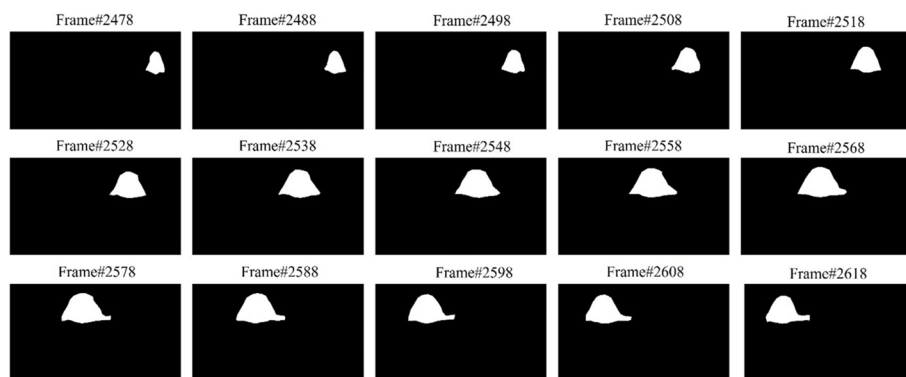
The training dataset is limited to only one sequence of 200 video frames per sample. Elaborately, only 200 video frames per train bogie are applied for training the model along with the defect-induced sequences. Our train video dataset consists of 7 video sequences, out of which six are normal and one is with defective bogie parts respectively. The total training video frames applied are  $7 \times 200 = 1.4K$ . Consequently, the remaining video frames are used for testing the trained model. The two networks in B-MHAC are trained and tested separately due to different hyperparameter initializations. The weight and bias initializations for B-MHAC has been through zero mean unit variance Gaussian distribution function. The learning rate in UNet was fixed across all datasets as 0.000001. This high learning rate enables the UNet to learn slowly over the entire object range. The bogie object masks were created using the annotation tool, ImageJ. The UNet is trained on stochastic gradient descent (sgd) optimizer and dice loss function. Specifically, the dice loss defines the overlap between the predicted and ground truth samples. To standardize the training process, the UNet across all datasets were trained for 100 epochs.

Consequently, the multi head classifier uses a dynamic learning rate initialized at 0.0001 which reduces by 10% when the error becomes constant for 10 continuous epochs. The momentum factor is 0.8. Here, we used the Adam optimizer for weight adaptations and cross entropy loss function for error calculations. The output of the classifier is a probability distribution function with maximum probability pointing towards the predicted class label. Additionally, inferencing on the test video sequences





**Fig. 8** UNet segmentation output for bogie part Axel across frames



**Fig. 9** Ground truth (GT) masks of the axle l in the dataset B-1

is accomplished by mapping the bounding box locations from the annotating data. The biggest advantage of the B-MHAC lies in the tiny training set that is sufficient for achieving robust performance over the entire test samples.

## Results and discussion

The proposed segmentation followed by multi head recognition of train bogie parts from high-speed video frames is being experimented with multiple datasets and variational hyperparameter combinations of the network during training. Subsequently, the results of the experiments were validated against the previous models on different test inputs. The following subsections provide a detailed analysis of the results obtained on multiple datasets.

### Quantitative validation of B-MHAC

First, we show the output of the UNet segmentation module on bogie train video sequences. Second, we present the three outcomes of the multi head classifier to show the confidence of the trained model in identifying a bogie object during inferencing. Figure 8 shows the results of the UNet on the axle bogie part. Simultaneously, the segmented axle bogie part is juxtaposed with ground truth sequences in Fig. 9.

The figures show only 15 frames of the video sequence in data B-1 when the train is moving from right to left of the screen. Subsequently, the results obtained for all other



presence of the segmentation module and the multi head attention network. The multi head attention network takes input from three sets of bogie parts at different time steps and generalizes on the location of the objects in the continuous video sequence. This has guaranteed greater accurate mapping of bounding box information onto the video sequence.

Consequently, the effectiveness of B-MHAC bogie part identification model is to be ascertained by comparing the results against popular image object detection models such as SSD, R-CNN, Fast R-CNN, Faster R-CNN, and our previous method with different Yolo versions. The visual results are presented in Fig. 11 on B-4 dataset. The proposed method outperformed other models due to the presence of multi head attention network that was learned in time steps on bogie object deformations.

This type of learning involves instances of both spatial and temporal information for classification making it robust to object deformations in the video sequences of moving trains. Finally, the B-MHAC is tested for defective parts identification on modified video sequences. The video frames with defective parts are fabricated with two defects on the spring suspension and binding screw.

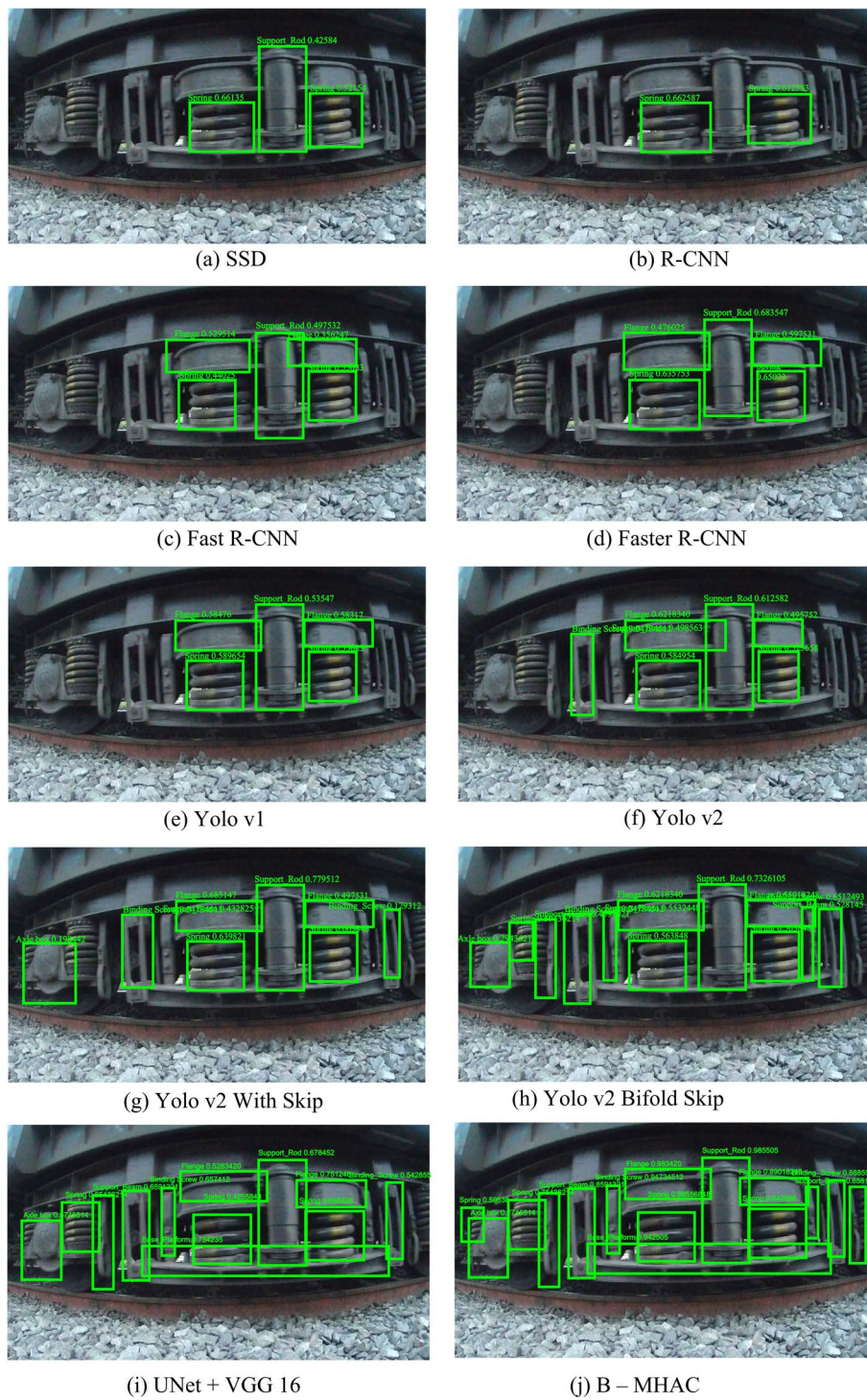
These defective part frames are induced into the video sequence, and the model was trained from scratch to identify defects by using the existing hyperparameters from the previous training. The results are projected onto the video sequence with a red bounding box for defective parts as shown in Fig. 12. The ability to identify defective parts by the proposed B-MHAC is found to be impressive. This is due to fact that the bogie part is segmented, and it passes through an attention span of multiple time steps which gives the network to learn distinct features across classes. Subsequently, the next subsection highlights the qualitative results on all the datasets with the calculated parameters as indicated above.

#### **Qualitative evaluation of B-MHAC model**

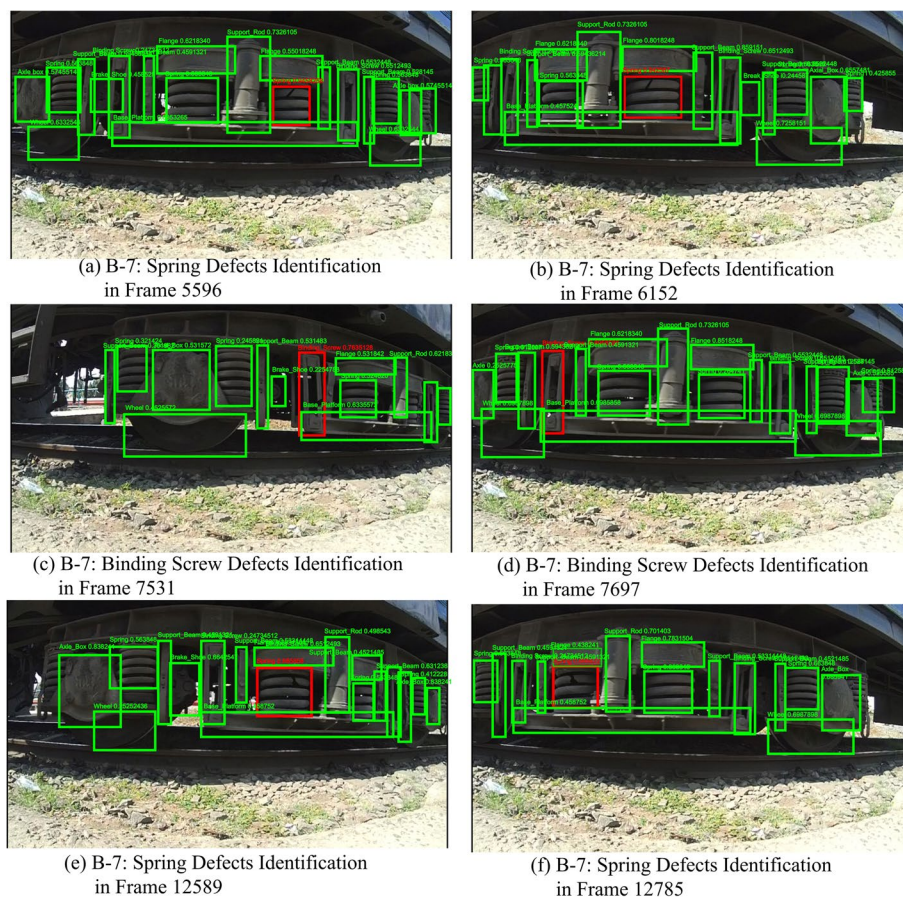
This subsection evaluates the proposed B-MHAC deep learning model on the six TRSE datasets. The IoU is calculated only for segmented bogie parts with UNet, and the remaining represents the classifier performance. The results are tabulated in Table 1. The values are averaged over the entire test sample. The average IoU across all datasets and bogie parts is 0.9162. This shows that the difference in predicted bogie segments and the GT has been narrowed extensively. We found a lower IoU for parts that are positioned at the end of the frames than that are in the middle. Additionally, the camera angle and the light intensities during recording also influenced the lower IoU scores on the datasets B-3 and B-6, respectively. Consequently, the average recognition mAP is 0.90115 across all datasets. Critical analysis showed that the bogie parts such as wheels and spring suspensions have recorded the lowest mAP values across all datasets. However, their scores were better than the previous models as shown in Table 2.

The models in Table 2 are trained from scratch on all datasets by keeping the hyperparameters constant. The other two parameters mFI and mNI indicate the B-MHAC failure to identify a part correctly and does not identify at all in the video frame. These two parameters are important in understanding the reason for the failure of the B-MHAC model. These parametric comparisons are presented in Tables 3 and 4. Analysis of these tables showcases that the bogie parts in the neighborhood of the camera focal length





**Fig. 11** Comparison of approaches for bogie part identification with the proposed B-MHAC (Zoom for better visibility). **a** B-7: spring defects identification in frame 5596. **a** SSD. **b** R-CNN. **c** Fast R-CNN. **d** faster R-CNN. **e** Yolo v2 with skip. **h** Yolo v2 bifold skip. **i** UNet + VGG 16. **j** B-MHAC



**Fig. 12** Defective parts identification through inferencing on trained B-MHAC (Zoom for better visibility). **a** B-7: spring defects identification in frame 5596. **b** B-7: spring defects identification in frame 6152. **c** B-7: binding screw defects identification in frame 7531. **d** B-7: binding screw defects identification in frame 7697. **e** B-7: spring defects identification in frame 12589. **f** B-7: spring defects identification in frame 12785

**Table 1** Performance of B-MHAC on TRSE video datasets

Datasets	IoU	mAP	mFI	mNI
B-1	0.9952	0.9639	0.1225	0.142
B-2	0.9057	0.8972	0.2004	0.2222
B-3	0.8987	0.883	0.2554	0.2365
B-4	0.9189	0.9503	0.1509	0.1489
B-5	0.9125	0.9177	0.1838	0.2053
B-6	0.8665	0.7948	0.4325	0.342
Average scores	0.91625	0.90115	0.22425	0.21615

have better identification potential than those that are away from it. In practice, it becomes extremely rigid to adjust the camera sensor position with respect to the moving train. Despite the above constraints, the B-MHAC has shown robust performance in instances where the camera sensor is randomly positioned. Overall, the B-MHAC has shown capabilities to sense bogie parts with exceptionally high accuracy when compared

**Table 2** Evaluation of B-MHAC against state-of-the-art methods on performance parameter mAP

Baseline methods/ datasets	SSD	R-CNN	Fast R-CNN	Faster R-CNN	Yolo v1	Yolo v2	Yolo v2 with skip	Yolo v2 bifold skip	B-MHAC
B-1	0.6843	0.6952	0.6856	0.7125	0.7752	0.7856	0.8152	0.9214	0.9587
B-2	0.6239	0.6531	0.6598	0.6859	0.7431	0.7658	0.7895	0.8152	0.9025
B-3	0.6151	0.6194	0.6252	0.6657	0.6894	0.7252	0.7594	0.8047	0.8956
B-4	0.6547	0.6773	0.6654	0.6913	0.6973	0.7754	0.7973	0.8478	0.9385
B-5	0.5955	0.6115	0.6175	0.6323	0.6615	0.7025	0.7415	0.8523	0.9122
B-6	0.5759	0.5936	0.5948	0.6189	0.6436	0.6868	0.7236	0.7321	0.8473
Average mAP	0.6249	0.6416	0.6413	0.6677	0.7016	0.7402	0.7710	0.8289	0.9091

**Table 3** Evaluation of the proposed method using mFI

Baseline methods/ datasets	SSD	R-CNN	Fast R-CNN	Faster R-CNN	Yolo v1	Yolo v2	Yolo v2 with skip	Yolo v2 bifold skip	B-MHAC
B-1	0.4215	0.4125	0.3785	0.3329	0.2882	0.2663	0.2156	0.1752	0.1124
B-2	0.4862	0.4598	0.4296	0.3889	0.3389	0.3025	0.2856	0.2531	0.1853
B-3	0.4621	0.4479	0.4129	0.3609	0.3268	0.2939	0.2556	0.2365	0.1722
B-4	0.4468	0.4352	0.4017	0.3569	0.3075	0.2701	0.2356	0.2036	0.1486
B-5	0.5374	0.5206	0.5251	0.5249	0.5161	0.5177	0.5056	0.4852	0.1672
B-6	0.5827	0.5933	0.5873	0.5789	0.5654	0.5215	0.5206	0.4952	0.2379
Average mFI	0.48945	0.4782	0.4558	0.4239	0.3904	0.362	0.3364	0.3081	0.1706

**Table 4** Evaluation of the proposed method using mNI

Baseline methods/ datasets	SSD	R-CNN	Fast R-CNN	Faster R-CNN	Yolo v1	Yolo v2	Yolo v2 with skip	Yolo v2 bifold skip	B-MHAC
D-1	0.5563	0.5125	0.4569	0.4236	0.3896	0.3456	0.3179	0.1856	0.1243
D-2	0.5936	0.5469	0.4856	0.4598	0.4189	0.3823	0.3495	0.2658	0.1975
D-3	0.6044	0.6093	0.5908	0.5815	0.5517	0.5355	0.5231	0.3856	0.2235
D-4	0.5459	0.4781	0.4282	0.3874	0.3603	0.3089	0.2863	0.1925	0.1385
D-5	0.5017	0.5037	0.4895	0.4312	0.3931	0.3622	0.3147	0.2489	0.1596
D-6	0.6271	0.6149	0.6121	0.6088	0.5924	0.5894	0.5515	0.4023	0.2578
Average mNI	0.5715	0.5442	0.5105	0.482	0.451	0.42	0.39	0.2801	0.1835

to other models. This is due to its multiple networks used for segmentation and recognition simultaneously.

#### Defect detection through B-MHAC and comparison

The primary objective of TRSE is to identify defective parts during transit. Therefore, any automated TRSE algorithm should have capabilities to detect defective parts in the vicinity of the normal bogie parts. This experiment is performed to test the ability



**Table 5** Experimental results showing defect identification abilities of TRSE automation models with mAP as the performance indicator

Baseline methods/ parameters	SSD	R-CNN	Fast R-CNN	Faster R-CNN	Yolo v1	Yolo v2	Yolo v2 with skip	Yolo v2 bifold skip	B-MHAC
mPA	0.4852	0.5325	0.5289	0.5475	0.6125	0.6589	0.6895	0.8745	0.9135
mFI	0.5987	0.5847	0.5245	0.5125	0.4528	0.4753	0.4236	0.2698	0.2258
mNI	0.5463	0.5126	0.5247	0.5169	0.4863	0.4236	0.4198	0.2891	0.2122

**Table 6** B-MHAC vs similar models on mAP

Baseline methods/ datasets	Block matching [1]	Active contours [2]	Shape prior active contours [3]	Shape invariance active contours [4]	Region- based active contours [5]	Unified active contour model [6]	Yolo v2 bifold skip	B-MHAC
B-1	0.8258	0.8525	0.8965	0.9145	0.9369	0.9522	0.9214	0.9627
B-2	0.7963	0.8256	0.8698	0.8963	0.9245	0.9289	0.8547	0.8963
B-3	0.7058	0.7256	0.7485	0.7458	0.7698	0.7852	0.7523	0.8425
B-4	0.8025	0.8266	0.8785	0.8989	0.9299	0.9369	0.9078	0.9457
B-5	0.7989	0.8158	0.8698	0.8858	0.9158	0.9195	0.8752	0.9025
B-6	0.6854	0.7125	0.7258	0.7458	0.7698	0.7896	0.7321	0.8147

of the algorithms to determine defective parts. Accordingly, the one set of training samples was selected as defective parts. In this work, only two defects were induced manually on the spring suspension and binding screw. A total of 200 frames were created with the two defects and were inducted into the video sequence of B-1. These are called broken part defects where the width and location of the cut are varied every 20 frames. The testing is performed with a 4000-frame video where 40 continuous frames were inducted into the original B-1 dataset at 5 randomly selected locations. The results of the experiment are shown in Table 5. Markedly, the proposed method shows robust defect identification capabilities over other methods by taking advantage of the multi head attention network. However, the model also suffers from inconsistency in defect dimensions which have gone undetected in the video sequence.

#### B-MHAC vs similar works

Previous works largely focused on the segmentation process of the bogie parts from the TRSE video sequences. These models aim to segment the bogie parts with precision rather than generating discriminative features for classification. Different from these approaches, we added an extra deep learning classifier at the end of segmentation processes for the recognition of bogie parts. This experiment will provide an insight into the behavior of the B-MHAC model over the existing models. Instead of including different CNN architectures along with the segmentation module, we applied our multi head attention classifier to the segmented outputs of these methods. The training the testing processes were in line with the original B-MHAC model. The results obtained are presented in Table 6. Only mAP was computed for a single

training run of these models. Apparently, the learning-based models have performed exceedingly better than the instance-based methods. Here, exclusively active contours have been used for segmentation of bogie parts with prior knowledge about the bogie part characteristics. Although it is evident that the active contours have been shown to possess superior segmentation quality, they have poor generalization capabilities on the test inputs. Hence, on video sequences with different camera angles, these models have performed weakly.

## Conclusions

An attempt has been made to apply deep learning approaches to automate TRSE. Initially, high-speed video sequences were recorded, and the dataset is created with high sparsity and resolution. A hybrid segmentation-classification method has been proposed to simultaneously segment and classify train bogie parts from video sequences. Contrasting the regular CNN models, we propose a multi stream multi head bogie part classifier (B-MHAC) on the segmented parts. Through extensive experimentation, it has been found that the proposed method resulted in an average recognition of 90.11%. The success of B-MHAC is credited to the attention mechanism at multiple time steps in the video sequence that helped the classifier to generalize better on the bogie part deformations on the running trains during recording. Furthermore, the approach has allowed for an automated interface environment where the TRSE can be performed remotely with high accuracy.

## Abbreviations

TRSE	Train rolling stock examination
UNET	U-shaped convolution network
CNN	Convolutional neural network
DBPI	Deep bogie part inspector
BPAS	Bogie part assert score
IR	Indian Railways
GPS	Global Positioning System
KRATES	Konkan Railways Automated Train Examination System
RGB	Red-green-blue
TGV	Train a Grande Vitesse
AC	Active contour
DL	Deep learning
GPU	Graphics processing unit
VGG	Visual Geometry Group
B-UNet	Bogie – U shaped convolutional neural network
HOG	Histogram of oriented gradients
BEC	Binary cross-entropy
FL	Focal loss
DL	Dice loss
B	MHACBogie — multi head attention classifier
IoU	Intersection over-union
Map	Mean Average Precision
Mfimean	False identification
mNI	Mean non-identification

## Acknowledgements

We thank the Indian Railways staff at Guntur for their expertise and assistance throughout all aspects of our study and for their help in data collection. We also thank the management of KLEF deemed to be university in helping us in all possible ways to accomplish this work.

## Authors' contributions

The author PVVK has conceptualized, validated, drafted, and edited the manuscript. KK has developed the methodology and the underlying code for the project. The manuscript was written by KK. Finally, video data collection, visualizations, and supervision were conducted by ChRP. The authors read and approved the final manuscript.

**Funding**

No funding was received to assist with the preparation of this manuscript.

**Availability of data and materials**

The datasets generated during and/or analyzed during the current study are not publicly available due to the memory constraints which are in excess of 47 GB but are available from the corresponding author on reasonable request in the form of 1-k video frames per train.

**Declarations****Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

Received: 21 April 2022 Accepted: 26 July 2022

Published online: 13 August 2022

**References**

- Kishore PVV, Prasad CR (2017) Computer vision based train rolling stock examination. *Optik* 132:427–444
- Kishore PVV, Prasad CR (2015) Train rolling stock segmentation with morphological differential gradient active contours. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, pp 1174–1178
- Sasikala N, Kishore PVV, Anil Kumar D, Prasad C (2019) Localized region based active contours with a weakly supervised shape image for inhomogeneous video segmentation of train bogie parts in building an automated train rolling examination. *Multimed Tools Appl* 78(11):14917–14946
- Sasikala N, Kishore PVV, Prasad CR, Kiran Kumar E, Anil Kumar D, Kumar MTK, Prasad MVD (2018) Unifying boundary, region, shape into level sets for touching object segmentation in train rolling stock high speed video. *IEEE Access* 6:70368–70377
- Mohan KK, Prasad CR, Kishore PVV (2021) Yolo v2 with bifold skip: a deep learning model for video based real time train bogie part identification and defect detection. *J Eng Sci Technol* 16(3):2166–2190
- Sasikala N, Kishore PVV (2020) Train bogie part recognition with multi-object multi-template matching adaptive algorithm. *J King Saud Univ Comput Inform Sci* 32(5):608–617
- Krishnamohan K, Prasad CR, Kishore PVV (2020) Successive texture and shape based active contours for train bogie part segmentation in rolling stock videos. *Int J Adv Comput Sci Appl* 11(6):589–598
- Chan TF, Vese LA (2001) Active contours without edges. *IEEE Transact Image Process* 10(2):266–277
- Lankton S, Tannenbaum A (2008) Localizing region-based active contours. *IEEE Transact Image Process* 17(11):2029–2039
- Tian B, Li L, Yansheng Q, Yan L (2017) Video object detection for tractability with deep learning method. In: 2017 Fifth International Conference on Advanced Cloud and Big Data (CBD). IEEE, pp 397–401
- Mandal M, Kumar LK, Saran MS (2020) MotionRec: a unified deep framework for moving object recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 2734–2743
- Ding X, Luo Y, Li Q, Cheng Y, Cai G, Munnoch R, Xue D, Qingying Y, Zheng X, Wang B (2018) Prior knowledge-based deep learning method for indoor object recognition and application. *Syst Sci Control Eng* 6(1):249–257
- Bi F, Ma X, Chen W, Fang W, Chen H, Li J, Assefa B (2019) Review on video object tracking based on deep learning. *J New Media* 1(2):63
- Ran X, Chen H, Zhu X, Liu Z, Chen J (2018) Deepdecision: a mobile deep learning framework for edge video analytics. In: IEEE INFOCOM 2018–IEEE Conference on Computer Communications. IEEE, pp 1421–1429
- Liu P, Qi B, Banerjee S (2018) Edgeeye: an edge service framework for real-time intelligent video analytics. In: Proceedings of the 1st international workshop on edge systems, analytics and networking, pp 1–6
- Olatunji IE, Cheng C-H (2019) Video analytics for visual surveillance and applications: an overview and survey. *Mach Learn Paradigms* 1:475–515
- Lee Y-H, Kim Y (2020) Comparison of CNN and YOLO for object detection. *J Semiconduct Display Technol* 19(1):85–92
- Schabert EJ, Hawley JA, Hopkins WG, Blum H (1999) High reliability of performance of well-trained rowers on a rowing ergometer. *J Sports Sci* 17(8):627–632
- Das NK, Das CK, Mozumder R, Bhowmik JC (2009) Satellite based train monitoring system. *J Electr Eng* 36(2):35–38
- Cacchiani V, Caprara A, Galli L, Kroon L, Maróti G, Toth P (2012) Railway rolling stock planning: robustness against large disruptions. *Transp Sci* 46(2):217–232
- Liu H, Li J, Song X, Seneviratne LD, Althoefer K (2011) Rolling indentation probe for tissue abnormality identification during minimally invasive surgery. *IEEE Transact Robot* 27(3):450–460
- Ashwin T, Ashok S (2014) Automation of rolling stock examination. In: 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies. IEEE, pp 260–263

23. Hart JM, Resendiz E, Freid B, Sawadisavi S, Barkan CPL, Ahuja N (2008) Machine vision using multi-spectral imaging for undercarriage inspection of railroad equipment. In: Proceedings of the 8th world congress on railway research, Seoul, Korea, vol. 18
24. Jarzebowicz L, Judek S (2014) 3D machine vision system for inspection of contact strips in railway vehicle current collectors. In: 2014 International Conference on Applied Electronics. IEEE, pp 139–144
25. Kazanskiy NL, Popov SB (2015) Integrated design technology for computer vision systems in railway transportation. *Pattern Recognit Image Anal* 25(2):215–219
26. Hwang J, Park H-Y, Kim W-Y (2010) Thickness measuring method by image processing for lining-type brake of rolling stock. In: 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content. IEEE, pp 284–286
27. Villar, Christopher M., Steven C. Orrell, II John Anthony Nagle. "System and method for inspecting railroad track." U.S. Patent 7,616,329, issued November 10, 2009.
28. Do NT, Gül M, Nafari SF (2020) Continuous evaluation of track modulus from a moving railcar using ANN-based techniques. *Vibration* 3(2):149–161
29. Lu H, Wang J, Shi H, Zhang D (2018) On-track experiments on the ride comforts of an articulated railway vehicle. In: Proceedings of the Asia-Pacific Conference on Intelligent Medical 2018 & International Conference on Transportation and Traffic Engineering 2018, pp 50–53
30. Meymand SZ, Keylin A, Ahmadian M (2016) A survey of wheel–rail contact models for rail vehicles. *Veh Syst Dyn* 54(3):386–428
31. Marques F, Magalhães H, Pombo J, Ambrósio J, Flores P (2020) A three-dimensional approach for contact detection between realistic wheel and rail surfaces for improved railway dynamic analysis. *Mech Mach Theory* 149:103825
32. Shams S, Platania R, Lee K, Park S-J (2017) Evaluation of deep learning frameworks over different HPC architectures. In: 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, pp 1389–1396
33. Sam SM, Kamardin K, Sjarif NNA, Mohamed N (2019) Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3. *Proc Comput Sci* 161:475–483
34. Qassim H, Verma A, Feinzimer D (2018) Compressed residual-VGG16 CNN model for big data places image recognition. In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, pp 169–175
35. Deshpande A, Estrela VV, Patavardhan P (2021) The DCT-CNN-ResNet50 architecture to classify brain tumors with super-resolution, convolutional neural network, and the ResNet50. *Neurosci Inform* 1(4):100013
36. Cheng D, Meng G, Cheng G, Pan C (2016) SeNet: Structured edge network for sea–land segmentation. *IEEE Geosci Remote Sensing Lett* 14(2):247–251
37. Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li L-J, Fei-Fei L, Yuille A, Huang J, Murphy K (2018) Progressive neural architecture search. In: Proceedings of the European conference on computer vision (ECCV), pp 19–34
38. Tian Y, Yang G, Wang Z, Wang H, Li E, Liang Z (2019) Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Comput Electron Agric* 157:417–426
39. Susanto Y, Livingstone AG, Ng BC, Cambria E (2020) The hourglass model revisited. *IEEE Intell Syst* 35(5):96–102
40. Li X, Chen H, Qi X, Dou Q, Chi-Wing F, Heng P-A (2018) H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transact Med Imaging* 37(12):2663–2674
41. Abdollahi A, Pradhan B, Alamri A (2020) VNet: an end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. *IEEE Access* 8:179424–179436
42. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transact Pattern Anal Mach Intell* 39(12):2481–2495
43. Kan M, Shan S, Chang H, Chen X (2014) Stacked progressive auto-encoders (spae) for face recognition across poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1883–1890
44. Thomas E, Pawan SJ, Kumar S, Anmol Horo S, Niyas SV, Kesavadas C, Rajan J (2020) Multi-res-attention UNet: a CNN model for the segmentation of focal cortical dysplasia lesions from magnetic resonance images. *IEEE J Biomed Health Inform* 25(5):1724–1734
45. Das A (2022) Adaptive UNet-based lung segmentation and ensemble learning with CNN-based deep features for automated COVID-19 diagnosis. *Multimed Tools Appl* 81(4):5407–5441
46. Sun Z, Huang S, Wei H-R, Dai X-y, Chen J (2020) Generating diverse translation by manipulating multi-head attention. *Proc AAAI Conf Artif Intell* 34(05):8976–8983
47. Dong J, Wang N, Fang H, Qunfang H, Zhang C, Ma B, Ma D, Haobang H (2022) Innovative method for pavement multiple damages segmentation and measurement by the Road-Seg-CapsNet of feature fusion. *Constr Build Mater* 324:126719
48. Ma D, Fang H, Wang N, Zhang C, Dong J, Hu H Automatic detection and counting system for pavement cracks based on PCGAN and YOLO-MF. In: IEEE transactions on intelligent transportation systems. <https://doi.org/10.1109/TITS.2022.3161960>
49. Xu Q, Liu Y, Li X, Yang Z, Wang J, Sbert M, Scopigno R (2014) Browsing and exploration of video sequences: a new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence. *Inf Sci* 278:736–756

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.